ATTORNEY DOCKET: R-341894

#13

Substitute Specification
for

CLONING, EXPRESSION AND CHARACTERIZATION OF THE SPG4 GENE RESPONSIBLE FOR THE MOST COMMON FORM OF AUTOSOMAL DOMINANT SPASTIC PARAPLEGIA

Inventors:
Jean Weissenbach, Jamilé Hazan

CERTIFICATE OF MAILING BY EXPRESS MAIL:

"Express Mail" Mailing Label No. **EL713287403US.**

I hereby certify that this paper and/or fee is being deposited with the United States Postal Service "EXPRESS MAIL POST OFFICE TO ADDRESSEE" service under 37 C.F.R. 1.10 on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231

Signature

05/02/2001
Date of Deposit

Joseph Krieger
Typed/printed name of person signing

CLONING, EXPRESSION AND CHARACTERIZATION OF THE SPG4 GENE RESPONSIBLE FOR THE MOST COMMON FORM OF AUTOSOMAL DOMINANT SPASTIC PARAPLEGIA.

BACKGROUND OF THE INVENTION

Field of Invention

The invention relates to the identification and characterization of the SPG4 gene encoding spastin, which is responsible for the most common form of autosomal dominant hereditary spastic paraplegia (HSP), to the cloning and characterization of its cDNA, and also to the corresponding polypeptides. The invention also relates to vectors, to transformed cells and to transgenic animals, and also to diagnostic methods and kits and to methods for selecting a chemical or biochemical compound capable of interacting directly or indirectly with a polypeptide according to the invention.

BACKGROUND OF THE INVENTION

Hereditary spastic paraplegias (HSPs) are degenerative disorders of the central nervous system, characterized by bilateral and progessive spasticity of the lower limbs. They reveal themselves clinically through difficulties in walking possibly evolving into total paralysis of both legs. The physiopathology of this set of diseases is, to date, relatively undocumented; however, anatomopathological data make it possible to conclude that the attack is limited to the pyramidal tracts responsible for voluntary motricity in the spinal cord (Reid, 1997). Various clinical and genetic forms of HSP exist. The so-called "pure" HSPs, which correspond to isolated spasticity of the lower limbs, are clinically distinguished from the "complex" HSPs, for which the spasticity of the legs is associated with other clinical signs of neurological or non-neurological type (Bruyn et al., 1991). From a genetic point of view, the HSPs can be transmitted according to the autosomal dominant (AD-HSP), autosomal recessive (AR-HSP) or X-linked (X-HSP) mode. The "pure" form of HSP, which is most commonly transmitted according to the autosomal dominant mode, remains the most frequent (approximately 80% of HSPs) (Reid, 1997). The incidence of HSPs, which remains difficult to estimate because of rare epidemiological studies and the considerable clinical variability, varies from 0.9 : 100 000 in Denmark, 3 to 9.6 : 100 000 in certain regions of Spain (Polo et al., 1991) or 14 : 100 000 in Norway (Skre, 1974) (approximately 3 : 100 000 in France).

In addition to this great clinical variability, which is observed not only between various families but also between various affected members of the same family, the HSPs are also characterized by considerable genetic heterogeneity. In the case of AD-HSPs, four loci have been identified, to date, on chromosomes 14 (locus SPG3) (Hazan et al., 1993), 2 (locus SPG4) (Hazan et al., 1994; Hentali et al., 1994), 15

(locus SPG6) (Fink et al., 1995) and 8 (locus SPG8) (Hedera et al., 1999). The study of a large number of families exhibiting an AD-HSP has shown that the gene carried by chromosome 2 is a main locus of this form of the disease, found in 40 to 50% of the families analyzed (The Hereditary Spastic Paraplegia Working Group, 1996; Durr et al., 1996). An anticipation phenomenon was observed in some locus SPG4-linked HSP families; this phenomenon has, subsequently, been associated with the expansion of a (CAG)n repeat demonstrated in 6 Danish families (Nielsen et al., 1997) using the RED (for Rapid Expansion Detection) technique. It has, however, never been possible to confirm this expansion in any of the families tested by this method or by the systematic search for sequences of (CAG)n type in physical maps composed of YAC (for Yeast Artificial Chromosome) or BAC (for Bacterial Artificial Chromosome) clones (Hazan et al., Genomics, 60 (3), 309-19, 1999).

To date, three genes responsible for two forms of X-HSP and one form of AR-HSP have been identified. Mutations in the gene which encodes a neuron-specific cell adhesion molecule, L1-CAM (for L1 Cell Adhesion Molecule), and which is located at Xq28 (locus SPG1) cause a complex form of HSP (Jouet et al., 1994) in which the spasticity is associated with a mental handicap, whereas mutations in the PLP (for ProteoLipid Protein) gene located at Xq21 (locus SPG2), which encodes a constitutive molecule of the myelin layer, cause pure and complex forms of X-HSP (Saugier-Veber, P. et al., 1994). More recently, mutations in the gene located at 16q24.3 (locus SPG7), which encodes paraplegin, a mitochondrial ATPase of the AAA (for "ATPases Associated with diverse cellular Activities") protein family (Confalonieri et al., 1995), have been associated with complex and pure forms of AR-HSP (Casari et al., 1998).

Thus, there remains, today, a great need to identify and characterize the gene responsible for the most common form of AD-HSP. The identification of this gene should, in particular, allow, besides the possibility of a test for antenatal screening in the families concerned, a better understanding of some of the molecular mechanisms engendering these degenerations specific for nerve bundles of the spinal cord, or even make it possible to provide an elementary response regarding therapeutic treatment for the patients.

SUMMARY OF The INVENTION

This is precisely the subject of the present invention.

After having delimited the localization range between the D2S352 and D2S2347 genetic markers by studying recombination events in locus SPG4-linked HSP families, the inventors have established a contig of BACs covering a physical distance evaluated at approximately 1.5 Mb and have undertaken a positional cloning strategy based on

sequencing the SPG4 range in order to completely identify all the genes located in the candidate region. The analysis of the sequence of the two BACs, D (b336P14) and G (B763N4), has revealed the presence of a gene which is composed of 17 exons, extending over a distance of approximately 100 kb, and which exhibits homology with the genes encoding proteins of the AAA family. Comparison of the sequence of this gene between the healthy and affected individuals of AD-HSP families has made it possible to demonstrate various mutations in the patients.

A subject of the invention is thus the identification and characterization of the SPG4 (or SPAST) gene encoding a novel nuclear member of the AAA family, responsible for the most common form of AD-HSP.

In a first aspect, a subject of the present invention is a purified or isolated nucleic acid of the SPG4 gene, characterized in that it comprises at least 15 consecutive nucleotides, preferably 20, 25, 30, 35, 40, 45, 50, 75, 100 or 200 consecutive nucleotides, of a sequence chosen from the group comprising:

- the sequence SEQ ID No. 1, which is a genomic sequence of the human SPG4 gene;
- the nucleic acid sequences which are homologs or variants of the nucleic acid of sequence SEQ ID No. 1;
- the sequence which is complementary thereto; and
- the sequence of the corresponding RNA thereof.

The present invention relates, of course, to both the DNA and RNA sequences, and also the sequences which hybridize with them, as well as the corresponding double-stranded DNAs.

The terms "nucleic acid", "nucleic acid sequence" or "sequence of nucleic acid", "polynucleotide", "oligonucleotide", "polynucleotide sequence", and "nucleotide sequence", which will be used equally in the present description, will be intended to refer to both a double-stranded DNA, a single-stranded DNA and products of transcription of said DNAs, and/or an RNA fragment, said isolated natural, or synthetic fragments which may or may not include unnatural nucleotides, referring to a precise series of nucleotides, which may or may not be modified, making it possible to define a fragment or a region of a nucleic acid. The expression "natural isolated, or synthetic DNA and/or RNA fragment, which may or may not include unnatural nucleotides" is intended to mean a precise series of nucleotides, which may or may not be modified, making it possible to define a fragment, a segment or a region of a nucleic acid.

It should be understood that the present invention does not relate to the genomic nucleotide sequences in their natural chromosomal environment, i.e. in the

natural state. It involves sequences which have been isolated and/or purified, i.e. they have been removed directly or indirectly, for example by copying, their environment having been at least partially modified.

The term "homologous nucleic acid sequence" is intended to refer to the sequences which have, with respect to the reference nucleic acid sequence, certain modifications, such as in particular a deletion, a truncation, an extension, a chimeric fusion and/or a mutation, in particular a point mutation, and the nucleic acid sequence of which shows at least 80%, preferably 90% or 95%, identity after alignment, with the reference nucleic acid sequence.

For the purpose of the present invention, the term "percentage of identity" between two nucleic acid or amino acid sequences is intended to refer to a percentage of nucleotides or of amino acid residues which are identical between the two sequences to be compared, obtained after the best alignment, this percentage being purely statistical and the differences between the two sequences being distributed randomly and throughout their length. Sequence comparisons between two nucleic acid or amino acid sequences are traditionally carried out by comparing these sequences after having optimally aligned them, said comparison being carried out by segment or by "window of comparison" in order to identify and compare local regions of sequence similarity. The optimal alignment of the sequences for comparison can be produced, besides manually, by means of the local homology algorithm of Smith and Waterman (1981) [Ad. App. Math. 2:482], by means of the local homology algorithm of Neddleman and Wunsch (1970) [J. Mol. Biol. 48:443], by means of the similarity search method of Pearson and Lipman (1988) [Proc. Natl. Acad. Sci. USA 85:2444], and by means of computer programs using these algorithms (GAP, BESTFIT, FASTA and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI, or with the BLAST N or BLAST P comparison programs).

The percentage of identity between two nucleic acid or amino acid sequences is determined by comparing these two optimally aligned sequences by window of comparison in which the region of the nucleic acid or amino acid sequence to be compared can comprise additions or deletions with respect to the reference sequence for optimal alignment between these two sequences. The percentage of identity is calculated by determining the number of identical positions for which the nucleotide or the amino acid residue is identical between the two sequences, dividing this number of identical positions by the total number of positions in the window of comparison and

multiplying the result obtained by 100 so as to obtain the percentage of identity between these two sequences.

For example, the BLAST program "BLAST 2 sequences" (Tatusova et al., "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol. Lett. 174:247-250), available on the site http://www.ncbi.nlm.nih.gov/gorf/bl2.html, may be used, the parameters used being those given by default (in particular for the parameters "open gap penalty" : 5, and "extension gap penalty" : 2; the matrix chosen being, for example, the "BLOSUM 62" matrix proposed by the program), the percentage of identity between the two sequences to be compared being calculated directly by the program.

It preferably involves sequences for which the complementary sequences are capable of hybridizing specifically with one of the sequences of the invention. Preferably, the specific or high stringency hybridization conditions will be such that they ensure at least 80%, preferably 90% or 95%, identity after alignment between one of the two sequences and the sequence which is complementary to the other.

Hybridization under high stringency conditions means that the temperature and ionic strength conditions are chosen such that they allow the hybridization between two complementary DNA fragments to be maintained. By way of illustration, high stringency conditions of the hybridization step for the purposes of defining the polynucleotide fragments described above are advantageously as follows.

The DNA-DNA or DNA-RNA hybridization is carried out in two steps: (1) prehybridization at 42°C for 3 hours in phosphate buffer (20 mM, pH 7.5) containing 5 x SSC (1 x SSC corresponds to a 0.15 M NaCl + 0.015 M sodium citrate solution), 50% of formamide, 7% of sodium dodecyl sulfate (SDS), 10 x Denhardt's, 5% of dextran sulfate and 1% of salmon sperm DNA; (2) actual hybridization for 20 hours at a temperature dependent on the size of the probe (i.e. 42°C for a probe of size > 100 nucleotides), followed by two 20-minute washes at 20°C in 2 x SSC + 2% SDS and one 20-minute wash at 20°C in 0.1 x SSC + 0.1% SDS. The final wash is carried out in 0.1 x SSC + 0.1% SDS for 30 minutes at 60°C for a probe of size > 100 nucleotides. The high stringency hybridization conditions described above for a polynucleotide of defined size will be adjusted by those skilled in the art for oligonucleotides of greater or smaller size, according to the teaching of Sambrook et al., 1989.

The term "nucleic acid sequence which is a variant" or "nucleic acid which is a variant" of a reference nucleic acid sequence will be intended to refer to the set of nucleic acid sequences corresponding to allelic variants, i.e. individual variations of the

reference nucleic acid sequence. These natural mutated sequences correspond to polymorphisms present in mammals, in particular in human beings, and in particular to polymorphisms which can cause a pathology to occur and/or to develop.

While the sequences according to the invention relate to normal sequences, they also relate to sequences which are mutated insofar as they include at least one point mutation, and preferably at most 10% of mutations, with respect to the normal sequence.

In particular, the variant nucleic acid sequences will comprise any sequence of at least 15 consecutive nucleotides, preferably 20, 25, 30, 50, 100 or 200 consecutive nucleotides, of a polymorphic sequence of the genomic sequence of the human SPG4 gene of sequence SEQ ID No. 1, and the nucleic acid sequence of which has, with respect to the sequence SEQ ID No. 1, at least one mutation corresponding in particular to a truncation, deletion, substitution and/or addition of an amino acid residue. In the present case, the variant nucleic acid sequences having at least one mutation will herein be linked to the pathologies of AD-HSP type linked to SPG4 locus.

Preferably, the present invention relates to the mutated nucleic acid sequences in which the mutations produce a modification of the amino acid sequence of the polypeptide encoded by the normal sequence.

The term "variant nucleic acid sequences" will also be intended to refer to any RNA or cDNA resulting from a mutation of a splice site of the genomic nucleic acid sequence SEQ ID No. 1.

Preferably, the invention relates to a purified or isolated nucleic acid of the SPG4 gene according to the invention, characterized in that it comprises a sequence chosen from the group comprising:

a) the sequence SEQ ID No. 1, the sequence SEQ ID No. 2, the sequence SEQ ID No. 72, the sequence SEQ ID No. 106 or the sequence of at least 15, preferably 20, 25, 30, 35, 40, 45, 50, 75, 100 or 200, consecutive nucleotides of the sequence SEQ ID No. 1, SEQ ID No. 2, SEQ ID No. 72 or SEQ ID No. 106;

b) the nucleic acid sequences which are homologs or variants of the sequences SEQ ID No. 1, SEQ ID No. 2, SEQ ID No. 72 or SEQ ID No. 106; and

c) the complementary sequence or the RNA sequence corresponding to the sequences as defined in a) and b),

preferably with the exception of the nucleic acid identified in the GenBank database under the accession number AB029006.

The nucleic acid the sequence of which is disclosed in the GenBank database under the accession number AB029006 corresponds to the sequence of one of the 100 cDNAs derived from a human brain mRNA library identified by the Kazusa DNA Research Institute in Japan (Kikuno et al., DNA Resarch, 6, 197-205, 1999).

Preferably, the invention relates to a purified or isolated nucleic acid according to the invention, characterized in that it comprises at least one sequence of at least 15 consecutive nucleotides, preferably 20, 25, 30, 50 or 75 consecutive nucleotides, of the nt 714-809, ends inclusive, fragment of the sequence SEQ ID No. 2, of the sequence complementary thereto or of the sequence of the corresponding RNA thereof.

The invention preferably relates to a purified or isolated nucleic acid according to the present invention, characterized in that it comprises a sequence chosen from the following group:

- the sequence SEQ ID No. 1;
- the sequence SEQ ID No. 2, which is the cDNA sequence encoding human spastin;
- the sequences SEQ ID No. 72 and SEQ ID No. 106, the sequence SEQ ID No. 72 representing the sequence of the incomplete cDNA encoding murine spastin represented in Figure 5, "mouse" line, and the SEQ ID No. 106 representing the complete sequence thereof;
- the nucleic acid sequences which are homologs or variants of the sequences SEQ ID No. 1, SEQ ID No. 2, SEQ ID No. 72 or SEQ ID No. 106;
- the sequence complementary thereto; and
- the sequence of the corresponding RNA thereof.

Preferably, the invention relates to a purified or isolated nucleic acid according to the invention, characterized in that it comprises at least one mutation which corresponds to a natural polymorphism in humans, in particular the position and nature of which are identified in Table 5.

The primers or probes, characterized in that they comprise a sequence of a nucleic acid according to the invention, also form part of the invention.

The present invention thus relates to the set of primers which can be deduced from the nucleotide sequences of the invention and which may make it possible to demonstrate said nucleotide sequences of the invention, in particular the mutated sequences, using in particular an amplification method such as the PCR method, or a related method.

The present invention also relates to the set of probes which can be deduced from the nucleotide sequences of the invention, in particular from the sequences

capable of hybridizing with them, and which may make it possible to demonstrate said nucleotide sequences, in particular to distinguish the normal sequences from the mutated sequences.

The present invention relates, in particular, to the probes or primers having sequences chosen from the sequences SEQ ID No. 4 to SEQ ID No. 71.

The invention also relates to the use of a nucleic acid sequence according to the invention as a probe or primer, for detecting, identifying, assaying or amplifying a nucleic acid sequence.

According to the invention, the polynucleotides which can be used as a probe or as a primer in processes for detecting, identifying, assaying or amplifying a nucleic acid sequence will have a minimum size of 15 bases, preferably of 20 bases, or better still of 25 to 30 bases.

The set of probes and primers according to the invention may be labeled directly or indirectly with a radioactive or nonradioactive compound, using methods well known to those skilled in the art, in order to obtain a detectable and/or quantifiable signal.

The nonlabeled polynucleotide sequences according to the invention can be used directly as a probe or primer.

The sequences are generally labeled so as to obtain sequences which can be used for many applications. The labeling of the primers or of the probes according to the invention is carried out with radioactive elements or with nonradioactive molecules.

Among the radioactive isotopes used, mention may be made of ^{32}P, ^{33}P, ^{35}S, ^{3}H or ^{125}I. The nonradioactive entities are selected from ligands, such as biotin, avidin or streptavidin, dioxygenin, haptens, colorants and luminescent agents, such as radioluminescent, chemiluminescent, bioluminescent, fluorescent or phosphorescent agents.

The polynucleotides according to the invention can thus be used as a primer and/or probe in processes using, in particular, the PCR (polymerase chain reaction) technique (Erlich, 1989; Innis et al., 1990, and Rolfs et al., 1991). This technique requires choosing pairs of oligonucleotide primers framing the fragment which must be amplified. Reference may, for example, be made to the technique described in American patent US No. 4,683,202. The amplified fragments can be identified, for example after agarose or polyacrylamide gel electrophoresis, or after a chromatographic technique such as gel filtration or ion exchange chromatography, and then sequenced. The specificity of amplification can be controlled using, as a primer,

the nucleotide sequences of polynucleotides of the invention and, as a matrix, plasmids containing these sequences or the derived amplification products. The amplified nucleotide fragments can be used as reagents in hybridization reactions in order to demonstrate the presence, in a biological sample, of a target nucleic acid having a sequence complementary to that of said amplified nucleotide fragments.

The invention is also directed toward the nucleic acids which can be obtained by amplification using primers according to the invention.

Other techniques for amplifying the target nucleic acid can be advantageously employed as an alternative to PCR (PCR-like), using pairs of primers having nucleotide sequences according to the invention. The term "PCR-like" will be intended to refer to all methods using direct or indirect reproductions of nucleic acid sequences, or in which the labeling systems have been amplified. These techniques are, of course, known. In general, they involve amplifying the DNA with a polymerase; when the sample of origin is an RNA, it is advisable to perform reverse transcription beforehand. There are, currently, a great many processes which enable this amplification, such as for example the SDA (Strand Displacement Amplification) technique (Walker et al., 1992), the TAS (Transcription-based Amplification System) technique described by Kwoh et al. in 1989, the 3SR (Self-Sustained Sequence Replication) technique described by Guatelli et al. in 1990, the NASBA (Nucleic Acid Sequence Based Amplification) technique described by Kievitis et al. in 1991, the TMA (Transcription Mediated Amplification) technique, the LCR (Ligase Chain Reaction) technique described by Landegren et al. in 1988 and improved by Barany et al. in 1991, which uses a heat-stable ligase, the RCR (Repair Chain Reaction) technique described by Segev in 1992, the CPR (Cycling Probe Reaction) technique described by Duck et al. in 1990, and the Q-beta-replicase amplification technique described by Miele et al. in 1983 and improved, in particular, by Chu et al. in 1986 and Lizardi et al. in 1988, and then by Burg et al., and also by Stone et al., in 1996.

When the target polynucleotide to be detected is an mRNA, use will advantageously be made, prior to carrying out an amplification reaction using the primers according to the invention or carrying out a detection process using the probes of the invention, of an enzyme of reverse transcriptase type in order to obtain a cDNA from the mRNA contained in the biological sample. The cDNA obtained will then serve as a target for the primers or probes used in the amplification or detection process according to the invention.

The probe hybridization technique can be carried out in diverse ways (Matthews et al., 1988). The most general method consists in immobilizing the nucleic acid extracted from the cells of various tissues or from cells in culture, on a support (such as nitrocellulose, nylon or polystyrene), and in incubating the immobilized target nucleic acid with the probe, under well defined conditions. After hybridization, the excess probe is eliminated and the hybrid molecules formed are detected using the appropriate method (measurement of the radioactivity, of the fluorescence or of the enzymatic activity linked to the probe).

According to another embodiment of the nucleic acid probes according to the invention, the latter can be used as a capture probe. In this case, a probe, termed "capture probe", is immobilized on a support and is used to capture, by specific hybridization, the target nucleic acid obtained from the biological sample to be tested, and the target nucleic acid is then detected using a second probe, termed "detection probe", labeled with an easily detectable element.

The splice acceptor or donor site sequences according to the present invention identified in Table 3 (sequences SEQ ID No. 74 to SEQ ID No. 105) also form part of the present invention.

In another aspect, the invention comprises a method for screening cDNA or genomic DNA libraries, or for cloning isolated genomic or cDNA encoding spastin, characterized in that it uses a nucleic acid sequence according to the invention.

Among these methods, mention may be made in particular of :

- the screening of cDNA libraries and the cloning of the isolated cDNAs (Sambrook et al., 1989; Suggs et al., 1981; Woo et al., 1979), using the nucleic acid sequences according to the invention;
- the screening of genomic libraries, for example of BACs (Chumakov et al., 1992; Chumakov et al., 1995), and, optionally, a genetic analysis by FISH (Cherif et al., 1990), using sequences according to the invention, enabling the isolation and chromosomal localization, and then the complete sequencing, of the SPG4 gene encoding spastin.

In particular, these methods according to the invention may be used for identifying and thus obtaining the genomic sequence or the cDNA of the SPG4 gene in other mammals, in particular mice.

These screening and/or cloning methods will comprise, in particular, a step of hybridization of a nucleic acid according to the invention with a nucleic acid contained in a genomic or cDNA library.

The invention also comprises a method for identifying the nucleic acid sequences which promote and/or regulate the expression of the SPG4 gene of sequence SEQ ID No. 1, characterized in that it uses a nucleic acid according to the invention.

The computer tools available to those skilled in the art enable them to easily identify, using the genomic nucleic acid sequences according to the invention, the promoter regulatory boxes required and sufficient for controlling gene expression, in particular the TATA, CCAAT and GC boxes, and also the stimulatory regulatory sequences ("enhancers"), or inhibitory regulatory sequences ("silencers"), which control, in CIS, the expression of the genes according to the invention; among these regulatory sequences, mention should be made of IRE, MRE and CRE.

The invention also relates to the methods for identifying mutations carried by the human SPG4 gene, in particular mutations responsible for autosomal dominant hereditary spastic paraplegia, characterized in that they use a nucleic acid sequence according to the invention.

These methods for identifying these mutations will, in particular, comprise the following steps: (i) isolation of the DNA from the biological sample to be analyzed, or production of a cDNA from the mRNA of the biological sample; (ii) specific amplification of the target DNA likely to have a mutation, using primers according to the invention; (iii) analysis of the amplification products, in particular the size and/or the sequence of the amplification products, with respect to a reference sequence.

The expression "methods for identifying a mutation according to the invention" is also intended to refer to a method which makes it possible to obtain the nucleic acid on which said mutation has been identified.

The promoter and/or regulatory sequences of the SPG4 gene according to the invention having mutations which may modify the expression of the corresponding protein also form part of the invention.

The nucleic acids characterized in that they can be obtained using one of the preceding methods according to the invention, or the nucleic acids capable of hybridizing, under high stringency conditions (homology of at least 80% between one of the two sequences and the sequence complementary to the other), with said nucleic acids, form part of the invention, especially the variant or homologous nucleic acids, in particular the nucleic acid sequences of allelic variants of the SPG4 gene of sequence SEQ ID No. 1 or of its cDNA of sequence SEQ ID No. 2, and also the genomic sequences of the homologous genes of other mammals such as mice.

In the present description, the term "Spg4" will be intended to refer to the mouse gene homologous to the human SPG4 gene.

The use of a nucleic acid sequence according to the invention as a probe or primer for screening a genomic library or a cDNA of course forms part of the subject of the present invention.

In another aspect, the invention comprises a purified or isolated polypeptide encoded by a nucleic acid according to the invention, preferably with the exception of the 584 amino acid peptide, the sequence of which is identified in the GenBank database under the accession number AB029006.

In the present description, the term "polypeptide" will be used to refer equally to a protein or a peptide.

Preferably, the present invention relates to a polypeptide according to the invention, characterized in that it comprises an amino acid sequence chosen from the following group:

- the sequence SEQ ID No. 3, corresponding to human spastin encoded by the sequence SEQ ID No. 2 of the cDNA of the human SPG4 gene;
- the sequence SEQ ID No. 73, corresponding to a fragment of murine spastin encoded by the sequence SEQ ID No. 72 of the incomplete cDNA of the mouse Spg4 gene, the sequence SEQ ID No. 73 is represented in Figure 4A, "SPAST_MOUSE" line;
- the sequence SEQ ID No. 107, corresponding to murine spastin encoded by the sequence SEQ ID No. 106 of the complete cDNA of the mouse Spg4 gene;
- the sequences of polypeptides which are homologs and variants of the polypeptide of sequence SEQ ID No. 3, SEQ ID No. 73 or SEQ ID No. 107; and
- the sequences of the fragments thereof of at least 8, 10, 15, 30 or 50 consecutive amino acids.

Also preferably, a subject of the invention is a polypeptide according to the invention, characterized in that it comprises an amino acid sequence chosen from the group comprising:

a) the sequence SEQ ID No. 3, the sequence SEQ ID No. 73, the sequence SEQ ID No. 107 or the sequence of at least 10 consecutive amino acids of one of these sequences; and

b) the sequences which are homologs or variants of the sequences SEQ ID No. 3, SEQ ID No. 73 or SEQ ID No. 107.

Also preferably, a subject of the invention is a polypeptide according to the invention, characterized in that it comprises the sequence of at least 8, preferably of at

least 10, 15, 20 or 30, consecutive amino acids of the sequence of the aa 197-228, ends inclusive, fragment of the sequence SEQ ID No. 3.

Also preferably, a subject of the invention is a polypeptide according to the invention, characterized in that it comprises an amino acid sequence chosen from the following group:

- the sequence SEQ ID No. 3, the sequence SEQ ID No. 73 and the sequence SEQ ID No. 107, which sequences carrying at least one of the mutations corresponding to a natural polymorphism in humans, in particular those the nature and location of which are identified in Table 5 hereinafter, or those which may be identified using the methods for identifying mutations of the SPG4 gene, according to the present invention; and
- the sequences of the fragments thereof of at least 8, 10, 15, 30 or 50 consecutive amino acids.

It should be understood that the invention does not relate to polypeptides in natural form, i.e. they are not taken in their environment. Specifically, the invention relates to the peptides which are obtained by purification from natural sources, or obtained by genetic recombination or by chemical synthesis, and which can therefore include unnatural amino acids. The production of a recombinant polypeptide, which can be carried out using one of the nucleotide sequences according to the invention, is particularly advantageous since it makes it possible to obtain an increased degree of purity of the desired polypeptide.

The term "homologous polypeptide" will be intended to refer to the polypeptides which have certain modifications with respect to the reference polypeptide, such as in particular one or more deletions or truncations, an extension, a chimeric fusion and/or one or more substitutions, and the amino acid sequence of which shows at least 80%, preferably 90% or 95%, identity after alignment, with the reference amino acid sequence.

The term "variant polypeptide" (or protein variant) will be intended to refer to the set of polypeptides encoded by the variant nucleic acid sequences as defined above.

In particular, the variant polypeptides will comprise any polypeptide which is encoded by the mutated genomic sequence of the SPG4 gene of sequence SEQ ID No. 1, and the amino acid sequence of which has at least one mutation corresponding in particular to a truncation, deletion, substitution and/or addition of amino acid residues with respect to the sequence SEQ ID No. 3. In the present case, the variant

polypeptides having at least one mutation will be linked to the pathologies of AD-HSP type.

The term "variant polypeptide" will also be intended to refer to any polypeptide resulting from mutation of a splice site in the genomic nucleic acid sequence SEQ ID No. 1.

The invention also comprises the cloning and/or expression vectors containing a nucleic acid sequence according to the invention.

The vectors according to the invention, characterized in that they include the elements which allow the expression and/or the secretion of said sequences in a host cell, or a cellular addressing sequence, also form part of the invention.

The vectors characterized in that they include a promoter and/or regulator sequence according to the invention also form part of the invention.

Said vectors will preferably include a promoter, translation initiation and termination signals, and also suitable regions for regulating the transcription. They should be able to be maintained stably in the cell and can, optionally, have particular signals which specify secretion of the translated protein.

These various control signals are chosen as a function of the host cell used. To this effect, the nucleic acid sequences according to the invention can be inserted into vectors which replicate autonomously in the host chosen, or vectors which integrate in the host chosen.

Among the systems which replicate autonomously, use will preferably be made, as a function of the host cell, of the systems of plasmid or viral type, the viral vectors possibly in particular being adenoviruses (Perricaudet et al., 1992), retroviruses, lentiviruses, poxviruses or herpesviruses (Epstein et al., 1992). Those skilled in the art know the technology which can be used for each of these systems.

When integration of the sequence into the chromosomes of the host cell is desired, use may be made, for example, of the systems of plasmid or viral type; such viruses will, for example, be retroviruses (Temin, 1986), or AAVs (Carter, 1993).

Among the nonviral vectors, preference is given to naked polynucleotides such as naked DNA or naked RNA according to the technique developed by the company VICAL, yeast artificial chromosomes (YAC) for expression in yeast, mouse artificial chromosomes (MAC) for expression in murine cells and, preferably, human artificial chromosomes (HAC) for expression in human cells.

Such vectors will be prepared according to the methods commonly used by those skilled in the art, and the clones resulting therefrom can be introduced into a

suitable host using standard methods, such as for example lipofection, electroporation or heat shock.

The invention also comprises the host cells, in particular the eukaryotic and prokaryotic cells, transformed with the vectors according to the invention, and also the transgenic animals, except humans, comprising one of said transformed cells according to the invention.

Among the cells which can be used for these purposes, mention may of course be made of bacterial cells (Olins and Lee, 1993), but also yeast cells (Buckholz, 1993), as well as animal cells, in particular cultures of mammalian cells (Edwards and Aruffo, 1993), and especially Chinese hamster ovary (CHO) cells, but also insect cells in which it is possible to use processes implementing baculoviruses, for example (Luckow, 1993). A preferred cellular host for expressing the proteins of the invention consists of CHO cells.

Among the mammals according to the invention, preference will be given to animals such as mice, rats or rabbits, expressing a polypeptide according to the invention.

Among the mammals according to the invention, preference will also be given to those comprising a transformed cell characterized in that the sequence of at least one of the two alleles of the SPG4 gene contains at least one of the mutations corresponding to a natural polymorphism in humans, in particular those the nature and location of which are identified in Table 5 hereinafter, or those which may be identified using the methods for identifying a mutation of the SPG4 gene, according to the present invention.

Among the mammals according to the invention, preference will also be given to animals such as mice, rats or rabbits, characterized in that the gene encoding spastin according to the invention is not functional or is knocked out.

Among the animal models more particularly advantageous herein, there are, in particular:

- the transgenic animals having, at least in one of their two allelic sequences of the SPG4 gene, at least one of the mutations the position and nature of which are identified in Table 5 or identified using a method according to the present invention These transgenic animals are obtained, for example, by homologous recombination on embryonic stem cells, transfer of these stem cells to embryos, selection of the chimeras affected in the reproductive lines, and growth of said chimeras;

- the transgenic animals (preferably mice) overexpressing the SPG4 gene into which one of said mutations according to the invention may be introduced. The mice are obtained, for example, by transfection of a copy of this gene under the control of a strong promoter which is ubiquitous in nature or selective for a tissue type, or after viral transcription;
- the transgenic animals (preferably mice) made deficient for the SPG4 gene according to the invention by inactivation using the LOXP/CRE recombinase system (Rohlmann et al., 1996) or any other system for inactivating the expression of this gene.

The cells and mammals according to the invention can be used in a method for producing a polypeptide according to the invention, as described below, and can also be used as a model for analysis and for DNA (genomic or cDNA) library screening.

The transformed cells or mammals as described above can thus be used as models in order to study the interactions between the polypeptides according to the invention, and chemical or protein compounds, which are involved directly or indirectly in the activities of the polypeptides according to the invention, this being in order to study the various mechanisms and interactions which come into play.

They can especially be used for selecting products which interact with the polypeptides according to the invention, in particular human spastin of sequence SEQ ID No. 3 or the variants thereof according to the invention, as a cofactor or as an inhibitor, in particular a competitive inhibitor, or which have agonist or antagonist activity for the activity of the polypeptides according to the invention. Preferably, said transformed cells or transgenic animals will be used as a model which, in particular, enables the selection of products which make it possible to combat the pathology linked to the SPG4 gene mentioned above.

The invention also relates to the use of a cell, of a mammal or of a polypeptide according to the invention for screening a chemical or biochemical compound which can interact directly or indirectly with the polypeptides according to the invention, and/or which is capable of modulating the expression or the activity of these polypeptides.

The invention also relates to the use of a nucleic acid sequence according to the invention for synthesizing recombinant polypeptides.

The method for producing a polypeptide of the invention in recombinant form is, itself, included in the present invention, and is characterized in that the transformed cells, in particular the cells or mammals of the present invention, are cultured under conditions which allow the expression of a recombinant polypeptide encoded by a

nucleic acid sequence according to the invention, and in that said recombinant polypeptide is recovered.

The recombinant polypeptides, characterized in that they can be obtained using said production method, also form part of the invention.

The recombinant polypeptides obtained as indicated above can be in both glycosylated and nonglycosylated form and may or may not have the natural tertiary structure.

These polypeptides can be produced based on the nucleic acid sequences defined above, according to the techniques for producing recombinant polypeptides known to those skilled in the art. In this case, the nucleic acid sequence used is placed under the control of signals which allow its expression in a cellular host.

An effective system for producing a recombinant polypeptide requires a vector and a host cell according to the invention.

These cells can be obtained by introducing into host cells a nucleotide sequences inserted into a vector as defined above, and then culturing said cells under conditions which allow the replication and/or expression of the transfected nucleotide sequence.

The processes for purifying a recombinant polypeptide which are used are known to those skilled in the art. The recombinant polypeptide can be purified from cell lyzates and extracts and/or from the culture medium supernatant, with methods used individually or in combination, such as fractionation, chromotography methods, immunoaffinity techniques using specific monoclonal or polyclonal antibodies, etc.

The polypeptides according to the present invention can be obtained by chemical synthesis, this using one of the many known peptide syntheses, for example the techniques which implement solid phases or techniques which use partial solid phases, by condensation of fragments or by conventional synthesis in solution.

The solid-phase synthesis technique is well known to those skilled in the art. See in particular Stewart et al. (1984) and Bodansky (1984).

The polypeptides which are obtained by chemical synthesis and which can include corresponding unnatural amino acids are also included in the invention.

The mono- or polyclonal antibodies or their fragments, chimeric antibodies or immunoconjugates, characterized in that they are capable of specifically recognizing a polypeptide according to the invention, form part of the invention.

Specific polyclonal antibodies can be obtained from a serum of an animal immunized against the polypeptides according to the invention, in particular produced

by genetic recombination or by peptide synthesis, according to conventional procedures.

The advantage of antibodies which specifically recognize certain polypeptides, variants or immunogenic fragments thereof, according to the invention, will in particular be noted.

The specific monoclonal antibodies can be obtained according to the conventional hybridoma culture method described by Köhler and Milstein, 1975.

The antibodies according to the invention are, for example, chimeric antibodies, humanized antibodies, or Fab or $F(ab')_2$ fragments. They can also be in the form of labeled antibodies or immunoconjugates in order to obtain a detectable and/or quantifiable signal.

The invention also relates to methods for detecting and/or purifying a polypeptide according to the invention, characterized in that they use an antibody according to the invention.

The invention also comprises purified polypeptides, characterized in that they are obtained using a method according to the invention.

Moreover, besides their use for purifying the polypeptides, the antibodies of the invention, in particular the monoclonal antibodies, can also be used for detecting these polypeptides in a biological sample.

They thus constitute a means of immunocytochemically or immunohistochemically analyzing the expression of the polypeptides according to the invention, in particular the polypeptide of sequence SEQ ID No. 3 or a variant thereof, on specific tissue sections, for example by immunofluorescence or gold labeling, or with an enzymatic immunoconjugates.

They may make it possible, in particular, to demonstrate abnormal expression of these polypeptides in the biological samples or tissues, which makes them useful for monitoring the progression of the disease and the molecular diagnosis.

More generally, the antibodies of the invention can be advantageously used in any situation in which the expression of a normal or mutated polypeptide according to the invention must be observed.

The methods for determining allelic variability, a mutation, a deletion, a loss of heterozygosity or any genetic abnormality of the SPG4 gene, according to the invention, characterized in that they use a nucleic acid sequence or an antibody according to the invention, also form part of the invention.

The present invention thus comprises a method for genotypic diagnosis of the pathology associated with the SPG4 gene, characterized in that a nucleic acid sequence according to the invention is used.

Preferably, the invention relates to a method for genotypic diagnosis of the disease associated with the presence of at least one mutation on a sequence of the SPG4 gene, using a biological sample from a patient, characterized in that it includes the following steps:

a) where appropriate, isolation of the genomic DNA from the biological sample to be analyzed, or production of cDNA from the RNA of the biological sample;

b) specific amplification of said DNA sequence of the SPG4 gene likely to contain a mutation, using primers according to the invention;

c) analysis of the amplification products obtained and comparison of their sequence with the corresponding normal sequence of the SPG4 gene.

The invention also comprises a method for diagnosing the disease associated with abnormal expression of a polypeptide encoded by the SPG4 gene, in particular the polypeptide of sequence SEQ ID No. 3, characterized in that one or more antibodies according to the invention is (are) brought into contact with the biological material to be tested, under conditions which allow the possible formation of specific immunological complexes between said polypeptide and said antibody or antibodies, and in that the immunological complexes possibly formed are detected and/or quantified.

These methods are, for example, directed toward the methods for diagnosis, in particular antenatal diagnosis, of AD-HSP associated with the presence of a mutation in the SPG4 gene, according to the invention, by determining, using a biological sample from the patient, the presence of mutations in at least one of the sequences described above. The nucleic acid sequences analyzed may equally be genomic DNA, cDNA or mRNA.

Nucleic acids or antibodies based on the present invention may also be used to enable positive diagnosis in a patient or presymptomatic diagnosis in an individual at risk, in particular an individual with a family history of the disease.

There are, of course, a great number of methods which make it possible to demonstrate a mutation in a gene with respect to the wild-type gene. They can essentially be divided into two main categories. The first type of method is that in which the presence of a mutation is detected by comparing the mutated sequence with the corresponding wild-type sequence, and the second type is that in which the presence

of the mutation is detected indirectly, for example through evidence of mismatches due to the presence of the mutation.

These methods can use the probes and primers of the present invention which have been described. They are generally purified nucleic acid hybridization sequences comprising at least 15 nucleotides, preferably 20, 25 or 30 nucleotides, characterized in that they can hybridize specifically with a nucleic acid sequence according to the invention.

Preferably, the specific hybridization conditions are such as those defined above or in the examples. The length of these nucleic acid hybridization sequences can range from 15, 20 or 30 to 200 nucleotides, particularly from 20 to 50 nucleotides.

Among the methods for determining allelic variability, a mutation, a deletion, a loss of heterozygocity or a genetic abnormality, preference is given to the methods comprising at least one so-called PCR (polymerase chain reaction) or PCR-like amplification step for the target sequence according to the invention likely to have an abnormality, using a pair of primers having nucleotide sequences according to the invention. The amplified products may be treated with a suitable restriction enzyme before carrying out the detection and assaying of the product targeted.

The mutations of the SPG4 gene according to the invention may be responsible for various modifications of the translation product thereof, these modifications possibly being used for a diagnostic approach. Specifically, the antigenicity modifications linked to these mutations may allow the development of specific antibodies. The mutated gene product can be distinguished using these methods. All these modifications can be employed in a diagnostic approach, using several well-known methods based on the use of mono- or polyclonal antibodies which recognize the normal polypeptide or mutated variants, such as for example by RIA or by ELISA.

Thus, a subject of the invention is also a kit or pack for diagnosis, in particular for diagnosing AD-HSP associated with the presence of a mutation in the SPG4 gene, according to the invention, characterized in that it comprises at least one compound chosen from the following group of compounds:

a) a nucleic acid, in particular as a primer or probe, according to the present invention; and

b) an antibody according to the invention.

In another aspect, the invention comprises a method for selecting a chemical or biochemical compound capable of preventing and/or treating AD-HSP associated with the SPG4 gene, characterized in that a nucleic acid sequence according to the

invention, a polypeptide according to the invention, a vector according to the invention, a cell according to the invention, a mammal according to the invention or an antibody according to the invention is used.

The methods for selecting chemical or biochemical compounds capable of interacting directly or indirectly with polypeptides according to the invention or with the nucleic acids according to the invention, and/or making it possible to modulate the expression or the activity of these polypeptides, characterized in that they comprise bringing a polypeptide according to the invention, a transformed cell according to the invention or a mammal according to the invention into contact with a candidate compound, and detecting a modification of the activity of said polypeptide, are also included in the invention.

For example, but without being limited thereto, mention may be made of a method for identifying molecules capable of interacting with a polypeptide according to the invention, using a bacterial or yeast two hybrid system such as the Matchmaker Two Hybrid System 2, according to the instructions of the manual which is supplied with the Matchmaker Two Hybrid System 2 (Catalog No. K1604-1, Clontech).

The nucleic acids encoding proteins which interact with the promoter and/or regulatory sequences of the SPG4 gene, according to the invention, can be screened and/or selected using a one hybrid system such as that described in the manual which is supplied with the Matchmaker One Hybrid System kit from Clontech (Catalog No. K1603-).

In other aspect, the invention comprises the use of a nucleic acid or of a polypeptide according to the invention, of a vector according to the invention, of a cell according to the invention or of a mammal according to the invention, for studying the expression or the activity of the SPG4 gene.

Other characteristics and advantages of the invention appear in the remainder of the description with the examples and figures, the legends of which are given hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

~~LEGENDS OF THE FIGURES~~

FIGURES 1A, 1B and 1C : Physical map of the SPG4 range and genomic organization of SPG4.

FIGURE 1A : The 1.5 Mb candidate region is delimited by the D2S352 and D2S2347 genetic markers indicated in bold characters. The position of the polymorphic markers and other STSs is indicated in standard characters, whereas the position of

the ESTs is indicated in italics. The BAC clones constituting the presequencing map are represented by rectangles, with the name shown above and the precise size of the clone, if it could be determined, shown below. The name of the BACs A, B, C, etc. is followed by brackets containing the name of the clone preceded by a "b" if the clone is derived from the BACs library CITB_978_SKB, or by a "B" if it originates from the library RPCI-11.

FIGURE 1B : Schematic representation of the SPG4 gene which overlaps BACs D (b336P14) and G (B563N4). The exons are shown as black rectangles with their name above.

FIGURE 1C : The five mutations identified in seven SPG4 locus-linked AD-HSP families are positioned in exons 7, 11 and 13 and in the splice acceptor site of intron 15.

FIGURE 2 : Nucleic acid and protein sequence of the SPG4 cDNA of spastin.

The 17 vertical bars with a number located below represent the junctions between the various exons. The ATG initiator codon is located at nt position 126-128 and the STOP codon for termination is located at nt position 1974-1976. Five of the mutations identified to date, including the loss of exon 16, are indicated in italics (nt 1210, nt 1468, nt 1520, nt 1620 and for the loss of exon 16: nt 1813-1853). The polyadenylation site is in italics and underlined. The putative nuclear localization signal (NLS), RGKKK, and also the three conserved domains predicted by the analysis in the ProDom database are located at aa positions 7-11 (NLS), 342-409 (domain 92), 411-509 (domain 179) and 512-599 (domain 6226), respectively. The four motifs predicted by the sequence comparison in the Prosite database are: two "leucine zipper" motifs at aa positions 50-78 and 508-529, the ATP binding site (or Walker A motif) at aa positions 382-389 and the "helix-loop-helix" dimerization domain at aa positions 478-486. The Walker A and B motifs, "GPPGNGKT" and "IIFIDE", and also the AAA minimum consensus [lacuna] are underlined.

FIGURES 3A and 3B : Characterization of a splice site mutation in the affected individuals of three SPG4 locus-linked AD-HPS families.

FIGURE 3A : PCR amplification of fragment IV of the SPG4 cDNA using lymphoblast cDNA: well M, size marker VII (Boehringer); well 1, unaffected member of family 2992; well 2, patient of family 2992; well 3, unaffected member of family 5330; well 4, patient of family 5330; well 5, patient of family 5226; well 6, negative control (human genomic DNA).

FIGURE 3B : Sequence graph for the mutation of the splice acceptor site of intron 15.

Genomic sequence of the control individual above and of a patient of family 2992 below. The asterisk at nt position 1813-4 indicates an A->C polymorphism which affects a nonconserved nucleotide of the splice acceptor site of intron 15 in the patient.

FIGURES 4A and 4B : Spastin homologies.

The identical residues are highlighted by shaded areas.

FIGURE 4A : Multiple alignment created by CLUSTAL W of eight proteins derived from various organisms and having strong sequence homology with human spastin and murine spastin (SEQ ID No. 73).

FIGURE 4B : Alignement by CLUSTAL W of the yeast metalloproteases AFG3, RCA1 and YME1, and of human plaraplegin and spastin.

FIGURE 5: Alignment by BLASTN of the nucleic acid sequences of the SPG4 cDNA and of its mouse ortholog Spg4 (SEQ ID No. 72). The polyadenylation site of the murine cDNA is underlined and in italics. The STOP codon is located at nt position 1515-1517 in the murine cDNA and at nt position 1974-1976 in the human cDNA.

FIGURES 6A, 6B and 6C : PCR analysis of the expression of SPG4 and of its murine ortholog Spg4.

FIGURE 6A : Collection of cDNA originating from multiple mouse tissues.

Well M, size marker V (Boehringer); well 1, heart, well 2, brain; well 3, spleen; well 4, lung; well 5, liver; well 6, skeletal muscle; well 7, kidney; well 8, testicle; well 9, E7 7-day embryo; well 10, E11 11-day embryo; well 11, E15 15-day embryo; well 12, E17 17-day embryo; well 13, negative control (mouse genomic DNA).

FIGURE 6B : Collection of cDNA originating from multiple human tissues.

Well M, size marker VII (Boehringer); well 1, brain; well 2, heart; well 3, kidney; well 4, liver; well 5, lung; well 6, pancreas; well 7, placenta; well 8, skeletal muscle, well 9, negative control (human genomic DNA); well 10, negative control (no DNA).

FIGURE 6C : Collection of cDNA originating from multiple human fetal tissues.

Well M, size marker VII (Boehringer); well 1, brain; well 2, heart; well 3, kidney; well 4, liver; well 5, lung; well 6, skeletal muscle; well 7, spleen; well 8, thymus; well 9, negative control (human genomic DNA); well 10, negative control (no DNA).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

EXAMPLES

Example 1: Materials and methods

1) Subcloning and sequencing of the candidate region

Twelve BACs originating from two human genomic libraries, CITB_978_SKB (sold by Research Genetics) and RPCI-11 (Osoegawa et al., 1998), and covering the SPG4 range, were selected to be sequenced (Hazan et al., Genomics, 60 (3), 309-19, 1999). 40 µg of the DNA of each BAC were partially digested with the CviJI restriction enzyme (CHIMERx) and separated by electrophoresis on 0.4% LMP agarose gel (FMC). DNA fractions, the sizes of which vary in the region of 3, 5 and 10 kb, were eluted with β-agarase (Biolabs) and ligated to a plasmid vector pBAM3, which had been digested with SmaI and dephosphorylated, beforehand, in a ratio of 1 × insert per 5 × vector. Electrocompetent E. coli DH10B bacteria (GIBCO-BRL) were transformed with the various ligations, by electroporation. Approximately 1 000 to 1 500 subclones per BAC (8 to 10 equivalent genomes), consisting of 20% of clones with inserts at 10 kb, 40% of clones with inserts at 5 kb and 40% of clones with inserts at 3 kb, were isolated. The ends of the inserts of these clones were sequenced on a LICOR 4200 automatic sequencer. For each BAC, the sequences were assembled into a backbone consisting of several contigs, using the Phred and Phrap programs. The holes between each contig were sequenced with labeled dideoxynucleotides on an ABI 377 sequencer (PE-Applied Biosystems). The exons contained in these sequence contigs were predicted with the GRAIL II, GENSCAN, FGENEH and Genie computer programs. The sequences were also compared in the EMBL and GenBank nucleic acid and protein databases, with the BLASTN and BLASTX programs. The determination of the promoter sequences was carried out using the TSSG and TSSW computer programs. The results of all these sequence analyses were visualized using the Genotator sequence annotation program.

2) cDNA cloning

The cDNA of the SPG4 gene was isolated through 5' and 3' RACE-PCR experiments on polyA+ RNAs of fetal brain, adult brain and adult liver, using the Marathon cDNA amplification kit (Clontech) according to the supplier's instructions. A first PCR followed by an internal PCR were carried out with various pairs of primers, the sequences of which are indicated in Table 1 hereinafter:

Table 1

Primers used for the RACE-PCRs and the cDNA amplifications

Primer	Sequence (5'-3')	5' position pair/PCR product size		
SPA_5RACE5	CGGAGCTCCTCTTGGCTGCCATG (SEQ ID No.4)	nt 405		
SPA_5RACE6	AGAAGCGCTGGCAGAGCCACACGAAG (SEQ ID No.5)	nt 372		
SPA_5RACE7	AAGGCGACCAAACGCAGCAGCGCGAAG (SEQ ID No.6)	nt 331		
SPA_3RACE1	AGGAGCAAGCTGTGGAATGGTATAAG (SEQ ID No.7)	nt 550		
SPA_3RACE2	TGGTTATGGCCAAGGACCGCTTACAAC (SEQ ID No.8)	nt 689		
SPA_3RACE3	CAAACGGACGTCTATAATGACAGTAC (SEQ ID No.9)	nt 747		
SPA_3RACE4	TTAGGAATGTGGACAGCAACCTTGC (SEQ ID No.10)	nt 1075		
SPA_3RACE5	CTTCTCTGAGGCCTGAGTTGTTCAC (SEQ ID No.11)	nt 1207		
SPA_3RACE6	TGCTAGAATGACTGATGGATACTCAGG (SEQ ID No.12)	nt 1736		
SPA_3RACE7	AGATGCAGCACTGGGTCCTATCCG (SEQ ID No.13)	nt 1787		
SPA_3RACE8	ATGAACGTCATCGGCTACAGAAACAG (SEQ ID No.14)	nt 2037		
SPA_Db	TAGCAGTGGCTGCCGCCGT (SEQ ID No.15)	nt 45	b+m	655 bp
SPA_Dm	AAGCGGTCCTTGGCCATAAC (SEQ ID No.16)	nt 700		
SPA_Dc	GGCGGCAGTGAGAGCTGTG (SEQ ID No.17)	nt 106	c+n	543 bp
SPA_Dn	CTAGCTCTTTCACACTGTTC (SEQ ID No.18)	nt 649		
SPA_Ad	AACAGGCCTTCGAGTACATC (SEQ ID No.19)	nt 487	d+n	746 bp
SPA_Am	CTGTGAACAACTCAGGCCTC (SEQ ID No.20)	nt 1233		
SPA_Ac	ATGAGAAAGCAGGACAGAAG (SEQ ID No.21)	nt 532		
SPA_An	TGCCAAGTCTTGACCAGC (SEQ ID No.22)	nt 1175		
SPA_Ba	CTACAACTGCTACTCGTAAG (SEQ ID No.23)	nt 1036	a+m	763 bp
SPA_Bm	CAGTGCTGCATCTTTTGCC (SEQ ID No.24)	nt 1799		
SPA_Bb	TAGGAATGTGGACAGCAACC (SEQ ID No.25)	nt 1076		
SPA_Bn	AAAGCTGTTAGGTCACTTCC (SEQ ID No.26)	nt 1780		
SPA_Ca	TGGAGATGACAGAGTACTTG (SEQ ID No.27)	nt 1550	a+m	766 bp
SPA_Cm	CTGGAATACTTTCATCTGC (SEQ ID No.28)	nt 2316		
SPA_Cb	ATGAGGCTGTTCTCAGGCG (SEQ ID No.29)	nt 1603		

The RACE-PCR products were cloned with the TA-cloning kit (Invitrogen) and the corresponding clones were sequenced on an ABI 377 (PE-Applied Biosystems). The sequence of the SPG4 transcript was varified by sequencing PCR products amplified from a cDNA population originating from the lymphoblasts of 6 healthy individuals.

3) Detection of mutations

The total RNAs were extracted from lymphoblast lines of one affected individual per family studied and of 6 control individuals, using the RNA PLUSR kit (bioprobe System). The cDNA synthesis was carried out on 500 ng to 1 µg of RNA, with 100 pmol of random hexameric primers (Pharmacia) and 200 units of Superscript II reverse transcriptase (Gibco BRL), under standard conditions. Four PCR amplifications, generating overlapping fragments which cover all of the SPG4 open reading frame, were carried out on the cDNAs of the patients and controls. Fragment I was amplified with the SPA_Db/SPA_Dm primers, and then by internal PCR with the SPA_Dc/SPA_Dn primers. Fragments II, III, and IV were amplified with the SPA_Ad/SPA_Am, SPA_Ba/SPA_Bm and SPA_Ca/SPA_Cm primers (cf. the sequences of these primers in Table 1), respectively. Each amplification was carried out in a total volume of 50 µl containing 4 µl of cDNA (~ 1/7th of the prep.), 20 pmol of each primer, 200 µM of dNTPs, 50 mM of KCl, 10 mM of Tris, pH 9, 1.5 mM $MgCl_2$, 0.1% of triton X-100, 0.01% of gelatin and 2.5 units of Taq polymerase (Cetus-PE). The PCR reactions were carried out according to the "hot start" process: the Taq polymerase is added at 92°C, after a first denaturation step of 5 min at 94°C. The samples are subsequently subjected to 35 cycles of denaturation (94°C for 40 sec), of hybridization (55°C for 50 sec, with the exception of fragment I: 58°C for 50 sec) and of elongation (72°C for 1 min), followed by a final elongation step (5 min at 72°C). The PCR products are sequenced on an ABI 377 automatic sequencer (PE-Applied Biosystems), with the SPA_Dc/SPA_Dn, SPA_Ac/SPA_An, SPA_Bb/SPA_Bn and SPA_Cb/SPA_Cm primers for fragments I, II, III and IV, respectively.

The mutations were also sought or confirmed by sequencing the 17 predicted exons of the SPG4 gene in the patients and controls. Each exon was amplified with the corresponding "a+m" pair of primers (cf. Table 2 hereinafter), with the exception of exon 1 (gSPAex1c/gSPAex1m), and exons 10, 11 and 12 which were co-amplified with the gSPAex10a/gSPAex12m and gSPAex11a/gSPAex12m pairs of primers.

Table 2

PCR primers for amplifying and sequencing the exons

Exon	Product size	PCR program	Primer	Sequence (5'-3') (SEQ ID Nos.; 30 to 71)
1	1048 bp	0	gSPAex1c	GTGAGCCGAACTGCACATTG
			gSPAex1m	CAAAGTCGACAGCTACAGTGC
			gSPAex1d	GGAACTGTAGTTGAGTGGGA
			gSPAex1n	AGATGAGGCTCCGACCTAC
2	624 bp	3	gSPAex2a	AATGCCACACTTGTAATCTC
			gSPAex2m	TGTGAATATATCATAATTTGGG
			gSPAex2b	TACAGCAGTTCTCATGATG
3	812 bp	1	gSPAex3a	GACCAAATTGGTGCATGCATG
			gSPAex3m	ACATTTCCAATACATCCCAC
4	379 bp	3	gSPAex4a	ATTTGTCATTTCACATGCAC
			gSPAex4m	TTAGAATGACTATACCTGAC
			gSPAex4n	TCAGGTTAAGTAAGACTC
5	830 bp	4	gSPAex5a	TTCCTATCTACCTAGTGAC
			gSPAex5m	TTTTATAGCAAGTTGCCCTG
			gSPAex5b	CCTATGAAGATCCTGGTAC
6	484 bp	3	gSPAex6a	TGTCATGATTCTAACAAGGG
			gSPAex6m	TCTATTTCACTCCTGACATG
7	420 bp	2	gSPAex7a	GTCATAGGGCTTAGGCTTC
			gSPAex7m	ATCATACTACCCACTTTTCC
8	647 bp	3	gSPAex8a	TGTTTGGGAAGATGCTACTG
			gSPAex8m	CTACTGAAGATAACGTACATG
9	1268 bp	1	gSPAex9a	CATTGATTGCCATGTATTGG
			gSPAex9m	AGAAGGCCAGAAATACTCAG
			gSPAex9b	GTACTTAAATCGGTAAATATGG
10	1061 bp	4	gSPAex10a	CTCAAGTCTTAGGAATGCAG
11			gSPAex10b	GCACTTAACCAGGCTGTATG
12	551 bp	3	gSPAex11a	CTCAGATGACTCACATAGC
			gSPAex12m	CTTTACTAGACTAATTCTCCTG

13	1361 bp	4	gSPAex13a	CAGATTCAAGAAGACAGATC
			gSPAex13m	GCAATAATTCACCACACTTG
			gSPAex13n	GGTAGTTCTTGTTTCTGCTC
14	985 bp	4	gSPAex14a	CAAGTGTGGTGAATTATTGC
			gSPAex14m	GAGCTGAAAAGTATTCAGC
			gSPAex14n	TGCAAAGGACATAGCCAGTG
15	1076 bp	1	gSPAex15a	AGCCTCTGGAGATAGTATGC
			gSPAex15m	CTAGAACAGGGGTCACAGTC
			gSPAex15n	TTGGACTTCTTAAACTTC
16	1404 bp	4	gSPAex16a	GCAGTATGCAAGAAATTGAAC
			gSPAex16m	GGCCTGTAATTTTCTTCTG
			gSPAex16b	GTACTGAATAGATACATGTAG
17	445 bp	3	gSPAex17a	GTGTAGCAGATCAACATAG
			gSPAex17m	CATCTTCAAGTTTGGTGCAC

Other than for exon 1, which is amplified using the Advantage GC genomic PCR kit (Clontech) according to the supplier's instructions, four slightly different PCR programs (1, 2, 3 and 4) were used to amplify the SPG4 exons (see Table 2). The amplifications were all carried out in a volume of 50 µl containing 100 ng of genomic DNA, 50 pmol of each primer, 250 µM pf dNTPs, 1X Takara buffer and 1 unit of Takara La Taq Taq polymerase (Shuzo Co.). The PCR reactions were carried out according to the "hot start" process: the Taq polymerase is added at 94°C, after a first denaturation step of 5 min at 96°C. The samples are subsequently subjected to 30 cycles of denaturation (94°C for 40 sec), of hybridization (prog. 1: 60°C for 50 sec; prog. 2: 58°C for 50 sec, prog. 3 and 4: 55°C for 50 sec) and of elongation (prog. 1 and 4: 72°C for 1 min, prog. 2 and 3: 72°C for 40 sec), followed by a final elongation step (10 min at 72°C). The sequencing of these PCR products was carried out on an ABI 377 sequencer (PE-Applied Biosystems), using either the PCR primers or the internal primers termed "b" and "n" (see Table 2).

4) Characterization of SPG4

The cDNA clones 977312 (EST AA560327) and 568234 (EST AA107866) derived from the mouse blastocyst and E8 embryo cDNA libraries, which both correspond to the murine ortholog of SPG4, were isolated using the IMAGE consortium

and sequenced in the laboratory on an ABI 377 sequencer (PE-Applied Biosystems). In order to analyze the expression profile of SPG4 and of its murine ortholog Spg4, the collections of cDNA from various fetal and adult human tissues, and also from mouse tissues (MTC panels, Clontech), were tested by PCR according to the supplier's protocol, with the SPA_Ca/SPA_Cm pair of primers for the human cDNAs and the SPA_Ca/spam (spam: 5'-ACCGAAGTCAAGAGCCTATC-3') pair for the mouse cDNAs. The PCR conditions are those used for amplifying SPG4 from lymphoblast line cDNA (cf. § Detection of mutations), except that these samples were subjected to 32 cycles for the cDNAs derived from adult human tissues and from mouse tissues, and to 28 cycles for the cDNAs derived from fetal tissues. The amplification products migrated by electrophoresis on 2% agarose gels.

5) Histological analysis of a muscle biopsy from a patient

The histological and histo-enzymatic analyses were carried out on a muscle biopsy from a patient derived from an SPG4 locus-linked family according to the standard techniques described in Casari et al., 1998.

6) Accession numbers in the public databases

The SPG4 (or SPAST) cDNA and the deduced protein sequence, GenBank/EMBL AJ246001; the incomplete Spg4 cDNA clone, GenBank/EMBL AJ246002; the SPG4 (or SPAST) gene, GenBank/EMBL AJ246003.

Example 2 : Analysis of the sequence of the SPG4 range

The analysis of the recombination events made it possible to reduce the SPG4 candidate region to a genetic range of 0 cM between the D2S352 and D2S2347 markers (19, 20). A presequencing map of the SPG4 range composed of 37 BACs was constructed (Hazan et al., in press in Genomics); the candidate region covers a physical distance of approximately of 1.5 Mb. Twelve overlapping BACs, stretching over the SPG4 region, with the exception of a single 4 kb hole between clones A and E, were selected to be sequenced (fig. 1A). Seven of these BACs (A, B, C, D, E, F and G), covering approximately 70% of the region of interest, have already been sequenced. The sequences of these 7 BACs were compared with those of the nucleic acid and protein databases, and analyzed with four exon prediction programs. These preliminary sequence analyses made it possible to reveal 14 potential transcription units, including three corresponding to the genes encoding xanthine dehydrogenase, steroid 5α-reductase 2 and a TGFβ-binding protein. Of the 14 genes detected by the sequence analysis, 9 had been previously identified in the EST (for "Expressed Sequence Tag") databases and located in the SPG4 range (Hazan et al., in press in

Genomics); the 5 remaining genes could only be identified by sequencing the candidate region. One of these 5 novel genes showed homology in 3' of its coding region, with the genes encoding the AAA protein family (Confalonieri et al., 1995). More thorough sequence analyses showed that this gene, named SPG4 (or SPAST), was composed of 17 exons and extended over a region of approximately 90 kb, covered by two adjacent BAC clones, D and G (cf. fig. 1B). The first three predicted exons of this gene were identified in BAC D, by two of the four exon prediction programs used, GRAIL II and GENSCAN; they show strong homology with a mouse blastocyst EST, AA560327. The last 14 exons are found in BAC G. The protein sequence deduced from exons 7 to 17 is significantly homologous to a subclass of the AAA family, which includes the Yta6p (Schnall et al., 1994), TBP6 (Schnall et al., 1994) and End 13 yeast proteins, and also the SKD1 mouse protein (Perier et al., 1994).

Of the four exon prediction programs FGENEH appears to be the most reliable and the most powerful, enabling detection of most of the genes of this chromosomal region at 2p21-p22. This observation also applies to the SPG4 gene, for which 15 exons could be demonstrated using this program, while only 4, 9 or 11 exons could be located using the Genie, GRAIL II and GENSCAN programs, respectively. The genomic organization of this gene (fig. 1B) could subsequently be confirmed by determining the sequence of the SPG4 cDNA. The intron/exon junctions are represented on table 3 hereinafter: the exon size ranges from 41 bp (exon 16) to 1.410 kb (exon 17), that of the introns ranging from 140 bp (intron 11) to 23.247 kb (intron 1).

Table 3

Intron/exon organization of the SPG4 gene

Exon/ intron	Exon size (bp)	Position on the cDNA	Splice acceptor site (SEQ ID No. 74 to 89)	Splice donor site (SEQ ID Nos. 90 to 105)	Intron size (bp)
1	540	1		TGAGAAAG/gtaactaggggctgg	23 247
2	87	541	atttttttattttaaag/CAGGACAG	AGGACAAG/gtaagattgtatttgt	1 943
3	84	628	aatttttttctttcag/GTGAACAG	ACTTCTAG/gtatcaattaatgtat	9 190
4	96	712	cttctctgttgcatag/AGAAGATG	CCAGTCAG/gtgggtttaggttaac	15 745
5	188	808	actttttccttgtcag/AAAGTGGA	CTCATAAG/gtattctgggacagta	876
6	134	996	ttttgtatcctttaag/GGTACTCC	GTGGACAA/gtaagttttgccatct	283
7	94	1 130	aggtcttgtttcttag/TGGAACAG	GGCCTGAG/gtaagaactttatatt	10 735
8	75	1 224	agtatatatttttttag/TTGTTCAC	CAATGCTG/gtaagggttctcttca	1 385
9	72	1 299	cttgtgatttttaaag/GCTAAAGC	CAAAATAC/gtgagtgctctgtttc	8 083
10	76	1 371	taatgctttgttttag/GTGGGAGA	TTTTATAG/gtaagaacatattttc	238
11	92	1 447	cttgtatttcctctag/ATGAAGTT	TTGATGGT/gtaagtgttgattatg	140
12	80	1 539	gatttttttgcttgtag/GTACAGTC	GTTCTCAG/gtagggagatttatat	4 715
13	43	1 619	ggatttttttttttag/GCGTTTCA	ATGAGGAG/gtatgtatctgtgttt	1 389
14	80	1 662	ttttaatatttttcag/ACAAGACT	CTTGCTAG/gtgagtaatttggatt	1 521
15	71	1 742	tccttcccttcctcag/AATGACTG	TATCCGAG/gtaggtatacaagagc	2 210
16	41	1 813	cttttatgttttacag/AACTAAAA	CCAGTGAG/gtatagtattttacaa	7 115
17	1 410	1 854	ctttttaaaaatctag/ATGAGAAA		

The sequences of the exons and introns are indicated in upper case and lower case, respectively.

Example 3 : Identification of the SPG4 cDNA

Several successive amplifications by 5' and 3' RACE-PCR were carried out on collections of adult liver and brain and fetal brain cDNA, in order to characterize the SPG4 transcript. All the 5' RACE-PCRs gave amplification products terminating at nt position 263 of the SPG4 cDNA (fig. 2), which was probably due to the rich GC content of the 5' region of the transcript (90% of GC in the 60 bp preceding nt position 263). Four overlapping PCR products, covering all of the coding region, were amplified from the cDNAs derived from the lymphoblasts of six control individuals, and entirely sequenced with the aim of verifying the sequence of the SPG4 transcript. Aligning the sequences of all the PCR and RACE-PCR products made it possible to reconstitute a 3263 bp sequence comprising a 1848 bp open reading frame preceded by a 125 bp untranslated 5' region (5' UTR for "5' UnTranslated Region") and followed by 1290 bp 3' UTR region including a polyadenylation site between nt positions 3227-3232, ~ 35 bp upstream of the polyA tail (fig. 2). Comparing the sequence of the SPG4 cDNA with the EST databanks made it possible to detect significant homology with 6 human ESTs, including EST N47973 which contains a more extended 3' noncoding region (+ 180 bp) comprising a second polyadenylation site. The translation initiation site was identified by the presence of a Kosak consensus sequence (CTGTGAatgA) defined as a "suitable context" for translation initiation given that a purine is located 3 nt upstream of the initiator ATG, itself preceded by a STOP codon. The 3263 bp cDNA sequence is identical to the transcribed sequence deduced from the 17 exons of the SPG4 gene. The analysis of the sequence of the 5' region using the TSSG and TSSW computer programs suggests the presence of a promoter sequence of the TATA box type located 43 bp upstream of nt position 1 of exon 1.

Example 4 : Mutations in the SPG4 gene

Heterozygous mutations were sought in the SPG4 cDNA originating from lymphoblasts of 14 patients derived from SPG4 locus-linked families (1 affected individual per family). Four overlapping PCR fragments, I, II, III and IV, covering the open reading frame of the SPG4 cDNA, were amplified and sequenced in the 14 patients, and also in 6 healthy control individuals. The agarose gel electrophoresis of PCR fragment IV showed three bands of equal intensity in 3 patients from families 2992, 5226 and 5330 originating from the same region of Switzerland, which would suggest a microdeletion or a mutation of a splice site; the two additional bands were not present in 2 healthy individuals derived from families 2992 and 5330 (fig. 3A). The genomic sequence of exon 16 revealed a heterozygous A->G mutation of the splice acceptor site (AG) of intron 15 in the affected

individuals of these three families (fig. 3B); this mutation engenders the loss of exon 16, followed by a reading frame shift in the abnormal transcript. None of the healthy members, including husbands and wives, carry this mutation of the splice site. The identification of the same mutation in all the affected members of these three Swiss families demonstrates the existence of a common ancestor, which had probably been suggested by the study of the haplotypes.

Three point mutations, 1210C->G, 1468G->A and 1620C->T, which introduced amino acid substitutions into the protein sequence (S362C, C448Y and R499C), were respectively revealed by sequencing PCR fragments III and IV in the affected individuals of families 624, 4014 and 618. These three substitutions all involve a cysteine residue, inducing the loss or insertion of a cysteine in the protein sequence. A 1 bp deletion, 1520delT, which creates the appearance of a STOP codon inducing a truncated protein composed of 465 amino acids (aa), was detected in the affected individuals of family A. None of the five mutations summarized in table 4 hereinafter was found in the control individuals tested, whether they belong to the healthy siblings or to the spouses of the seven families analyzed herein. These five mutations significantly affect the protein sequence in a very conserved domain, or AAA cassette (Beyer, 1997), which is composed of several protein motifs presumed to be responsible for the ATPase activity in all the members of the AAA family.

Table 4

Mutations in SPG4 in the patients suffering from AD-HSP

Family	Location	Mutation [a]	Amino acid change [b]	Consequence
624	exon 7	1 210 C ↦ G	S362C	missense
4 014	exon 11	1 468 G ↦ A	C448Y	missense
A	exon 11	1 520 delT	466STOPcodon	nonsense
618	exon 13	1 620 C ↦ T	R499C	missense
2 992	intron 15	1 813-2a ↦ g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift
5 226	intron 15	1 813-2a ↦ g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift
5 330	intron 15	1 813-2a ↦ g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift

[a] The nt positions refer to the sequence of the SPG4 cDNA.

[b] The aa positions refer to the spastin sequence.

The bases of the exons are indicated in upper case, those of the introns in lower case.

PTC+7 aa = "premature termination codon" at 7 aa downstream of exon 16.

In addition to these five mutations described above, searches for heterozygous mutations, carried out on patients suffering from AD-HSP derived from 36 other families, made it possible to reveal 34 other mutations which modified or were likely to modify the product of expression of the SPG4 gene.

The characteristics of these 34 other mutations are summarized in table 5 hereinafter, into which the first five mutations mentioned above have also been inserted.

Tabl 5

Mutations in SPG4 in the patients suffering from AD-HSP

Family	Location	Mutation [a]	Amino acid change [b]	Consequence
624	exon 7	1210 C ↦G	S362C	missense
6958	exon 8	1233 G ↦A	G370R	missense
214	exon 8	1267 T ↦G	F381C	missense
1002	exon 8	1283 T ↦G	N386K	missense
027	exon 8	1288 A ↦G	K388R	missense
019	exon 10	1401 C ↦G	L426V	missense
4014	exon 11	1468 G ↦A	C448Y	missense
148	exon 11	1504 G ↦T	R460L	missense
618	exon 13	1620 C ↦T	R499C	missense
636	exon 15	1788 G ↦A	D555N	missense
627	exon 15	1792 C ↦T	A556V	missense
2971	exon 3	702 C ↦T	Q193STOP	nonsense
3655	exon 5	873 A ↦T	K229STOP	nonsense
1010	exon 5	907 C ↦A	S261STOP	nonsense
3938	exon 5	932 C ↦G	Y269STOP	nonsense
6922	exon 10	1416 C ↦T	R431STOP	nonsense
616	exon 10	1416 C ↦T	R431STOP	nonsense
605	exon 15	1809 C ↦T	R562STOP	nonsense
030	exon 2	578-579insA	PTC + 2 aa	shift + nonsense
615	exon 5	852del11	PTC + 18 aa	shift + nonsense
042	exon 5	882-883insA	PTC + 12 aa	shift + nonsense
032	exon 5	906delT	PTC + 17 aa	shift + nonsense
189	exon 9	1299delG	PTC + 3 aa	shift + nonsense
3686	exon 9	1340del5	PTC + 35 aa	shift + nonsense
625	exon 9	1340del5	PTC + 35 aa	shift + nonsense
A	exon 11	1520delT	PTC + 7 aa	shift + nonsense
115	exon 12	1574delGG	PTC + 2 aa	shift + nonsense
3266	exon 13	1634del22	PTC + 18 aa	shift + nonsense
149	exon 14	1684-1685insTT	PTC + 9 aa	shift + nonsense
645	exon 14	1685del4	PTC + 7 aa	shift + nonsense
029	intron 4	808-2 a ↦g	?	
162	intron 6	1129+2 t ↦g	?	splice site mutation
125	intron 7	1223+1 g ↦t	?	splice site mutation
143	intron 8	1299+1 g ↦a	?	splice site mutation
1620	intron 11	1538+5 g ↦a	(PTC + 6 aa)	splice site mutation
1006	intron 11	1538+3 del4	?	loss of exon 11 + shift
1605	intron 13	1661+1 g ↦t	?	splice site mutation
1012	intron 13	1662-2 a ↦t	?	splice site mutation
1626	intron 15	1812+1 g ↦a	?	splice site mutation
2992	intron 15	1813-2 a ↦g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift
5226	intron 15	1813-2 a ↦g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift
5330	intron 15	1813-2 a ↦g	Δ aa564 ↦ aa576 (PTC+7 aa)	loss of exon 16 + shift
1611	intron 16	1813-2 a ↦g	?	splice site mutation
		1853+1 g ↦a		

[a] The nt positions refer to the sequence of the SPG4 cDNA. [b] The aa positions refer to the spastin sequence. The exon bases are indicated in upper case, those of the introns in lower case. PTC+n aa - "premature termination codon" at n amino acids downstream of the mutation.

Example 5 : Analysis of the protein sequence of spastin

The open reading frame of SPG4 encodes a 616 aa protein which we have named spastin and the molecular weight of which is approximately 67.2 kDaltons (kD). The comparison of this amino acid sequence in the protein databases, using the BLAST programs, made it possible to reveal a region of strong homology with several members of the AAA family, at the C-terminal end of spastin. The "typical" motifs of the AAA family, encompassed in the AAA cassette, are located between aa positions 342 and 599 (see fig. 2) according to the sequence comparisons in the ProDom and Prosite protein domain databases. The three conserved typical domains, including the Walker A and B motifs and also the minimum consensus motif of the AAA proteins are located in the AAA cassette at aa positions 382-389, 437-442 and 480-498, respectively, (fig. 2). The Walker A motif, "GPPGNGKT", also called p-loop, which corresponds to the ATP-binding domain, and the B motif, "IIFIDE", are very conserved among all the members of the AAA family, including spastin.

The comparison of the AAA cassettes present in 150 proteins of this ATPase family, derived from organisms which are very far apart in evolution made it possible to classify this set of proteins into several subgroups, as a function of the number of AAA cassettes identified (1 or 2) and of the sequence homologies between these various cassettes (Beyer, 1997). Among all the proteins of the AAA family, spastin shows stronger homology with a particular subclass of the AAAs, and more specifically with the following proteins, most of which were identified through the complete sequencing of the genome of the organism in question: two proteins of Caenorhabditis elegans, O16299 and Q18128; two subunits of the 26S proteasome of Saccharomyces cerevisiae, Yta6p (Q02845) and TBP6 (P40328) (Schnall et al., 1994); a subunit of the proteasome of Schizosaccharomyces pombe (O43078); the SAP1 (P39955) and END13 (P52917) proteins of S. cerevisiae and the murine SKD1 protein (P46467) (Perier et al., 1994). The multiple alignment of these 8 proteins with spastin is represented in fig. 4A. Of the 257 amino acids encompassing the AAA cassette (aa positions 342-599), spastin shows 52%, 51% and 50% sequence identity with the Yta6p (Q02845) yeast protein, the O16299 nematode protein and the TBP6 (P40328) yeast protein, respectively. Similar results were obtained by analyzing the protein sequence of spastin in the ProDom database, which showed the existence of three domains of homology (named 92, 179 and 6226, and corresponding to aa positions 342-409, 411-509 and 512-599) found in the putative subunits of the 26S proteasome of yeast. In addition, the members of this AAA subgroup most commonly contain motifs of the leucine-zipper type, two of which could be detected

in the protein sequence of spastin at aa positions 50-78 and 508-529, by analyzing the sequence in the Prosite database (see fig. 2). This analysis was also able to predict the presence of a dimerization motif of the helix-loop-helix type, located between aa positions 478 and 486.

The comparison of the protein sequence of spastin with those of the mitochondrial metalloproteases, such as the AFG3, RCA1 and YME1 yeast proteins, and also paraplegin, which is implicated in a rare form of AR-HSP, shows that the homology between these five members of the AAA family is limited to the 257aa region encompassing the AAA cassette (fig. 4B). In this region, the sequence identity between spastin and paraplegin is only 29%, whereas paraplegin and the AFG3 yeast protein are 57% identical over this same portion of the protein sequence. This sequence comparison suggests that spastin does not belong to the same AAA subgroup as paraplegin and other mitochondrial metalloproteases. In addition, the computer analysis of the spastin sequence using the PSORT II program, which makes it possible to predict the subcellular location of the proteins, appears to indicate that spastin is a nuclear protein. A possible nuclear localization signal (NLS), RGKKK, was revealed between aa positions 7 and 11, whereas no signal peptide characteristic of importation into mitochondria could be detected, unlike what had been observed for paraplegin.

Example 6 : Expression profiles for SPG4 and for its murine ortholog Spg4

The comparison of the nucleic acid sequence of SPG4 in the EST databanks made it possible to detect several human, murine and rat ESTs showing strong homology with SPG4. The mouse blastocyst and E8 embryo cDNA clones corresponding to two of the murine ESTs, AA560327 and AA107866, were obtained from the IMAGE consortium and entirely sequenced. The assembly of the sequences of these cDNA clones made it possible to reconstitute a 1689 bp consensus sequence including a 1514 bp incomplete open reading frame. The comparison between the human SPG4 cDNA and this mouse cDNA showed that the murine transcript lacks approximately 460 bp at the 5' end, including the translation initiation codon. The mouse open reading frame is followed by a 175 bp 3' noncoding region (3' UTR) containing a polyadenylation site located ~20 bp upstream of the polyA tail (fig. 5). The nucleic acid sequence of SPG4 and the protein sequence of human spastin show 89% (between nt positions 460 and 1982) and 96% (between aa positions 113 and 616) identity, respectively, with the mouse cDNA and deduced protein sequences. This considerable degree of homology makes it possible to affirm that this mouse transcript corresponds to the murine ortholog of SPG4, which was therefore named Spg4.

The hybridization of Northern blots comprising the mRNAs of various human and murine tissues (Clontech) with the SPG4 and Spg4 cDNA clones did not give any convincing results, except a very weak band corresponding to a 2.5 kb transcript in the mouse testicle after exposure for 10 days. Because of the low level of expression of this gene, the expression profiles for SPG4 and Spg4 were determined by PCR experiments on normalized collections of cDNA originating from various adult and fetal tissues (see fig. 6A to 6C). The murine Spg4 gene is expressed ubiquitously in the adult tissues of mice, and also from the E7 stage to the E17 stage of mouse embryos (fig. 6A). Higher expression of Spg4 was detected in the liver, skeletal muscle and testicles, and also at the E15 stage of embryos. The early expression of Spg4 during embryonic development was confirmed by the presence of ESTs originating from blastocyst, E8 embryo and embryonic carcinoma cDNA libraries in the public EST databanks. The human SPG4 gene is, itself, also expressed ubiquitously in adult (fig. 6B) and fetal (fig. 6C) tissues, with perhaps more marked expression in fetal brain.

Example 7 : No oxidative phosphorylation impairment in SPG4 locus-linked AD-HSP

In order to determine whether spastin mutations induced an oxidative phosphorylation (OXPHOS) impairment in mitochondria, in the same way as had been observed for paraplegin, a muscle biopsy was performed on a patient from one of the SPG4 locus-linked AD-HSP families. The morphological and histo-enzymatic analyses of this muscle biopsy did not reveal any muscle fibers of the RRF (for "ragged red fiber") type, characteristic of OXPHOS impairments in mitochondria. The fact that all the muscle fibers appear to be normal, and also the prediction of a nuclear localization for spastin, seem to indicate that SPG4 locus-linked AD-HSP is not a mitochondrial disease of the OXPHOS type, unlike SPG7 locus-linked AR-HSP.

Using a positional cloning approach based on sequencing a 1.5 Mb region, we have identified the SPG4 (or SPAST) gene responsible for the most common form of AD-HSP, previously located on chromosomal bands 2p21-p22. Thirty nine mutations which modify or are likely to modify the gene product, named spastin, could be detected in the affected individuals from forty one families with AD-HSP showing a link to the SPG4 locus. Spastin is a novel member of the AAA protein family, which appears to have a nuclear localization and which shows strong homology with the subunits of the 26S proteasome of yeast. Despite great homology restricted to a domain of 230 to 250 aa, termed AAA cassette, the many members of this protein family can participate in very varied cellular mechanisms, such as the transport of proteins in vesicles, cell cycle

regulation, organelle biogenesis, i.e. control of transcription, etc. However, all these cellular mechanisms involve the assembly, the functioning or the degradation of protein complexes, which suggest that the members of the AAA family are so-called "chaperon" proteins.

References

Barany, F., (1991), Proc. Natl. Acad. Sci. USA, 88, 189-193.

Beyer, A. Sequence analysis of the AAA protein family. Protein Sci. 6, 2043-2058 (1997).

Bodansky M., Principles of peptide synthesis, (1984).

Bruyn, R.P.M. & Scheltens, P.H. Hereditary spastic paraparesis (Strumpell-Lorrain) in Handbook of clinical neurology Vol. 15 (ed. de Jong, J.M.B.V.) 301-318 (Elsevier Science Publishers B.V., 1991).

Buckholz, R.G. Curr. Op. Biotechnology 4 : 538-542, 1993.

Burg, J.L. et al. (1996), Mol. and Cell. Probes, 10, 257-271.

Carter, B.J. Curr. Op. Biotechnology 3 : 533-539, 1993.

Casari, G. et al. Spastic paraplegia and OXPHOS impairment caused by mutations in Paraplegin, a nuclear-encoded mitochondrial metalloprotease. Cell 93, 973-983 (1998).

Cherif D., Julier, C., Delattre, O., Derré, J., Lathrop, G.M., and Berger, R. Proc. Natl. Acad. Sci. USA. 87 : 6639-6643, 1990.

Chu, B.C.F. et al. (1986), Nucleic Acids Res., 14, 5591-5603.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billault, A., Guasconi, G., Gervy, P., Le Chumakov, I.M., Rigault, P., Le Gall, I., et al. Nature 377 : 175-183, 1995.

Confalonieri, F. & Duguet, M. A 200-amino acid ATPase module in search of a basic function. BioEssays 17, 639-650 (1995).

Duck, P. et al. (1990), Biotechniques, 9, 142-147.

Durr, A. et al. Phenotype of autosomal spastic paraplegia linked to chromosome 2. Brain 119, 1487-1496 (1996).

Edwards, C.P., and Aruffo, A. Curr. Op. Biotechnology 4 : 558-563, 1993.

Epstein, A. Médecine/Sciences 8 : 902-911, 1992.

Erlich, H.A., (1989), New York : Stockton Press.

Guatelli J.C. et al. Proc. Natl. Acad. Sci. USA 87: 1874-1878, 1990 et al. Cell 85 : 281-290, 1996.

Fink, J.K. et al. Autosomal dominant familial spastic paraplegia : tight linkage to chromosome 15q. Am. J. Hum. Genet. 56, 188-192 (1995).

Hazan, J., Lamy, C., Melki, J., Munnich, A., de Recondo, J., & Weissenbach, J. Autosomal dominant familial spastic paraplegia is genetically heterogeneous and one locus maps to chromosome 14q. Nature Genet. 5, 163-167 (1993).

Hazan, J. et al. Linkage of a new locus for autosomal dominant familial spastic paraplegia to chromosome 2p. Hum. Mol. Genet. 3, 1569-1573 (1994).

Hedera, P. et al. Novel locus for autosomal dominant hereditary spastic paraplegia, on chromosome 8q. Am. J. Hum. Genet. 64, 563-569 (1999).

Heinzlef, O. et al. Mapping of a complicated familial spastic paraplegia to locus SPG4 on chromosome 2p. J. Med. Genet. 35, 89-93 (1998).

Hentati, A. et al. Linkage of a locus for autosomal dominant familial spastic paraplegia to chromosome 2p markers. Hum. Mol. Genet. 3, 1867-1871 (1994).

The Hereditary Spastic Paraplegia Working Group. Hereditary spastic paraplegia : advances in genetic research. Neurology 46, 1507-1514 (1996).

Innis, M.A. et al. (1990), Academic Press.

Jouet, M. et al. X-linked spastic paraplegia (SPG1), MASA syndrome and X-linked hydrocephalus result from mutations in the L1 gene. Nature Genet. 7, 402-407 (1994).

Kievitis, T. et al. (1991), J. Virol. Methods, 35, 273-286.

Köhler et Milstein. Nature 256, 495-497, 1975.

Kwoh, D.Y. et al. (1989), Proc. Natl. Acad. Sci. USA, 86, 1173-1177.

Landegren U., Kaiser R., Sanders J. & Hood L. Science 241 : 1077-1080, 1988.

Lizardi, P.M. et al. (1988), Bio/technology, 6, 1197-1202.

Luckow, V.A. (1993), Curr. Op. Biotechnology 4, 564-572.

Matthews, J.A. et al. (1988), Anal. Biochem., 169 : 1-25.

Miele, E.A. et al. (1983), J. Mol. Biol., 171 : 281-295.

Nielsen, J.E. et al. CAG repeat expansion in autosomal dominant pure spastic paraplegia linked to chromosome 2p21-24. Hum. Mol. Genet. 6, 1811-1816 (1997).

Olins, P.O., and Lee, S.C. Curr. Op. Biotechnology 4 : 520-525, 1993.

Osoegawa, K. et al. An improved approach for construction of bacterial artificial chromosome libraries. Genomics 52, 1-8 (1998).

Perier, F. et al. Identification of a novel mammalian member of the NSF/CDC48p/Pas1p/TBP-1 family through heterologous expression in yeast. FEBS lett. 351, 286-290 (1994).

Perricaudet, M., Stratford-Perricaudet, L. and Briand, P. La Recherche 23 : 471-473, 1992.

Polo, J.M., Calleja, J., Combarros, O. & Berciano, J. Hereditary ataxias and paraplegias in Cantabria, Spain. An epidemiological and clinical study. Brain 114, 855-866 (1991).

Reid, E. Pure hereditary spastic paraplegia. J. Med. Genet. 34, 499-503 (1997).

Rohlmann, A., Gotthardt, M., Willnow, T.E., Hammer, R.E., and Herz, J. Nature Biotech. 14 : 1562-1565, 1996.

Rolfs, A. et al. (1991), Berlin : Springer-Verlag.

Sambrook, J., Fritsch, E.F., and Maniatis, T. Molecular cloning : a laboratory manual.

Saugier-Veber, P. et al. X-linked spastic paraplegia and Pelizaeus-Merzbacher disease are allelic disorders at the proteolipid protein locus. Nature Genet. 6, 257-262 (1994).

Schnall, R. et al. Identification of a set of yeast genes coding for a novel family of putative ATPases with high similarity to constituents of the 26S protease complex. Yeast 10, 1141-1155 (1994).

Scott, W.K. et al. Locus heterogeneity, anticipation, and reduction of the chromosome 2p minimal candidate region in autosomal dominant familial spastic paraplegia. Neurogenetics 1, 95-102 (1997).

Sec. Ed. Cold Spring Harbor Lab., Cold Spring Harbor, New York.

Segev, D., (1992), Kessler C. Springer Verlag, Berlin, New-York, 197-205.

Skre, H. Hereditary spastic paraplegia in Western Norway. Clin. Genet. 6, 165-183 (1974).

Stone, B.B. et al. (1996). Mol. and Cell. Probes, 10 : 359-370.

Stewart J.M. et Yound J.D., solid phase peptides synthesis, Pierce Chem. Company, Rockford, 111, 2ème éd., (1984).

Suggs S.V., Wallace R.B., Hirose T., Kawashima E.H. and Itakura K. PNAS 78 : 6613-6617, 1981.

Temin, H.M. Retrovirus vectors for gene transfer. In Kucherlapati R., ed. Gene Transfer, New York, Plenum Press, 149-187, 1986.

Walker G.T., Fraiser M.S., Schram J.L., Little M.C., Nadeau J.G., & Malinowski D.P. Nucleic Acids Res. 20 : 1691-1696, 1992.

Werderlin, L. Hereditary ataxias. Occurence and clinical features. Acta Neurol. Scand. 73 (Suppl. 106) (1986).

Woo S.L.C. Methods Enzymol. 68 : 389, 1979.